Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Caracterización del trabajo infantil en el Perú 2019, usando árboles de decisión

TRABAJO DE INVESTIGACIÓN

Para optar el Grado Académico de Bachiller en Estadística

AUTOR

Keith CABALLERO RODRIGUEZ

ASESOR

Mg. Emma Norma CAMBILLO MOYANO

Lima, Perú

2021

Referencia bibliográfica

Caballero, K. (2021). *Caracterización del trabajo infantil en el Perú 2019, usando árboles de decisión*. [Trabajo de investigación de bachiller, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Hoja de metadatos complementarios

Código ORCID del autor	https://orcid.org/0000-0001-9116-5190
DNI o pasaporte del autor	71478831
Código ORCID del asesor	https://orcid.org/0000-0003-3173-9425
DNI o pasaporte del asesor	15377390
Grupo de investigación	Ciencias matemáticas para las ciencias de la vida https://vrip.unmsm.edu.pe/cmatvida/
Agencia financiadora	—
Ubicación geográfica donde se desarrolló la investigación	Lugar: Perú Coordenadas geográficas: -9.189967 -75.015152
Año o rango de años en que se realizó la investigación	2019
Disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

Universidad del Perú. Decana de América

FACULTAD DE CIENCIAS MATEMÁTICAS

ESCUELA PROFESIONAL DE ESTADÍSTICA

ACTA DE SUSTENTACIÓN DE TESIS EN LA MODALIDAD VIRTUAL PARA OBTENER EL GRADO DE BACHILLER EN ESTADISTICA

Siendo las 14:00 horas del 1 de marzo del 2021, en la Sala de Sesión Virtual de la Facultad de Ciencias Matemáticas, se reunieron los docentes designados como miembros del Jurado Evaluador:

Dr. Roger Pedro Norabuena Figueroa	(Presidente)
Mg. Roberto Fiestas Flores	(Miembro)
Mg. Emma Norma Cambillo Moyano	(Miembro Asesor)

Para la sustentación del Trabajo de investigación intitulado **"CARACTERIZACIÓN DEL TRABAJO INFANTIL EN EL PERÚ 2019, USANDO ÁRBOLES DE DECISIÓN",** presentado por el Sr. **KEITH CABALLERO RODRIGUEZ,** para obtener el Grado de Bachiller en Estadística.

Luego de la exposición del Trabajo de investigación, el Presidente invitó al expositor a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del Jurado Evaluador, el expositor mereció la **APROBACIÓN CON MENCIÓN HONROSA**, con un calificativo promedio de **DIECIOCHO (18)**.

A continuación, los miembros del Jurado Evaluador, dan manifiesto que el participante Sr. **KEITH CABALLERO RODRIGUEZ**, en virtud de haber aprobado la sustentación de su Trabajo de investigación, será propuesto para que se le otorgue el Grado de Bachiller en Estadística.

Siendo las 14:50 horas, se levantó la Sesión, firmando para constancia la presente Acta, en archivo pdf.

UNMSM Firmado digitalmente por HUAMAN GUTIERREZ Zoraida Judith FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 07.04.2021 10:49:06 -05:00

Dr. Roger Pedro Norabuena Figueroa
Presidente

Mg. Roberto Fiestas Flores
Miembro

Mg. Emma Norma Cambillo Moyano
Miembro Asesor

La Vicedecana Académica (e) de la Facultad de Ciencias Matemáticas, Mg. Zoraida Judith Huamán Gutiérrez, certifica virtualmente la participación del Jurado Evaluador, la estudiante, el acto de instalación y el inicio, desarrollo y término del acto académico de sustentación, dejando constancia en el acta respectiva.

Resumen

CARACTERIZACIÓN DEL TRABAJO INFANTIL EN EL PERÚ 2019, USANDO ÁRBOLES DE DECISIÓN

Keith Caballero Rodriguez

Febrero 2021

Asesora : Mg. Emma Norma Cambillo Moyano
Grado obtenido : Bachiller en Estadística

Objetivo: Caracterizar el trabajo infantil en el Perú en el año 2019

Metodología: El trabajo de investigación es no experimental, descriptivo y correlacional transversal. La fuente secundaria fue extraída de la ENAHO 2019 con la cual la muestra conformaba 29962 menores de 5 a 17 años con registros completos de 6,872,105 como población. Se realizó la prueba de chi cuadrado y estadísticos univariados como bivariados, además de aplicar el modelo de árboles de decisión CART encontrando las variables más importantes.

Resultados: En base al modelo de árboles de clasificación se encontró que los perfiles asociados los trabajadores infantiles son en su mayoría aquellos que poseen primaria incompleta, viven en hogares cuyo jefe de hogar realiza actividades de agricultura, silvicultura y pesca, localizados en las regiones de Áncash, Apurímac, Cajamarca y Huancavelica; se ubican en la costa norte, sierra y selva, un grupo menores de 10 a 11 años; en adición a menores de 9 años que viven en las regiones de Cusco y Pasco..

Conclusiones: Se obtuvo que la región del hogar de residencia del menor, la edad, el último nivel de estudios aprobado y la actividad realizada por el jefe(a) de hogar son los más importantes predictores para caracterizar el trabajo infantil, además de encontrar un modelo no paramétrico e interpretable con una sensibilidad 0.72 el cual representa un 72 % de trabajadores infantiles que fueron clasificados correctamente como trabajadores.

Palabras clave: Trabajo infantil**,** caracterización, árboles de decisión

Abstract

CHARACTERIZATION OF CHILD LABOR IN PERU 2019, USING DECISION TREES

Keith Caballero Rodriguez

February 2021

Advisor : Mg. Emma Norma Cambillo Moyano
Grade : Bachelor in Statistics

Objective: Characterize child labor in Peru in 2019

Methodology: The research work is non-experimental, descriptive and cross-correlational. The secondary source was extracted from the ENAHO 2019, with which the sample consisted of 29,962 minors between 5 and 17 years of age with complete records from 6,872,105 as population. The chi-square test and univariate and bivariate statistics were performed, in addition to applying the decision tree model (CART), finding the most important variables.

Results: It was found in that the profiles associated with child workers are mostly those who have incomplete primary school, live in households whose head of household carries out agricultural, forestry and fishing activities, located in the regions of Áncash, Apurímac, Cajamarca and Huancavelica; They are located on the north coast, mountains and jungle, a profile is under 12 years old and another is between 10 to 11 years old; In addition there are children under 9 years of age who live in the regions of Cusco and Pasco.

Conclusions: It was obtained that the region of the home of residence of the minor, age, the last approved level of education and the activity carried out by the head of the household are the most important predictors to characterize child labor, in addition to finding a model that is not parametric and interpretable with a sensitivity of 0.72 which represented 72% of child workers who were correctly classified as workers

Keywords: Child labor, characterization, decision trees

CARACTERIZACIÓN DEL TRABAJO INFANTIL EN EL PERÚ 2019, USANDO ÁRBOLES DE DECISIÓN

Keith Caballero Rodriguez

Informe presentado a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el grado de Bachiller en Estadística.

Aprobado por:

Dr. Roger Pedro Norabuena Figueroa Presidente	Mg. Roberto Fiestas Flores Miembro

Mg. Emma Norma Cambillo Moyano
Miembro Asesor

Lima – Perú

Febrero – 2021

FICHA CATALOGRÁFICA

Keith Caballero Rodriguez

CARACTERIZACIÓN DEL TRABAJO INFANTIL EN EL PERÚ 2019, USANDO ÁRBOLES DE DECISIÓN

Lima 2021

X (número de páginas de los preliminares), 40 (número de páginas del informe) p., 29.7 cm (UNMSM, Bachiller, Estadística, 2019).

Universidad Nacional Mayor de San Marcos
Facultad de Ciencias Matemáticas
Estadística
UNMSM / FCM

DEDICATORIA

A mis padres y hermanos los cuales han sido la clave para poder seguir alcanzando mis metas día a día, y el aprecio que sienten me permitieron llegar hasta el final de la carrera. A mis maestros quienes nunca desistieron al enseñarme.

ÍNDICE DE CONTENIDOS

INDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1. Introducción

El trabajo infantil realizado por niños, niñas y adolescentes (NNA) es un problema bastante amplio de tratarse, este a su vez es definido usualmente como la realización de tareas que priva a los menores de su niñez, su potencial y dignidad, y es perjudicial para su desarrollo físico y psicológico (Organismo Internacional del Trabajo – OIT,2020).

Aunque no existe aún una definición estadística homologada internacionalmente, las normas jurídicas internacionalmente proveen de flexibilidad a la hora de delimitar los umbrales para determinar que formas de trabajo son consideradas o se encuentran comprendidas como trabajo infantil, en general para la medición del trabajo infantil los países estructuran la definición considerando la edad del NNA y el tipo de actividades productivas que realizan (OIT, 2018). El incumplimiento de los derechos fundamentales de los NNA repercute negativa y continuamente a lo largo de sus vidas (Comisión Económica para América Latina y el Caribe - CEPAL,2009). Además, este problema tiene repercusiones en el gobierno, afectando la macroeconomía, las características propias del menor y su familia, políticas de gobiernos y la pobreza (OIT y CEPAL,2018). En el Perú se considera a un NNA en situación de trabajo infantil a los que se encuentran entre 5 a 11 años que realizan al menos una hora a la semana, una o más actividades económicas, en el marco de la frontera de la producción del Sistema de Cuentas Nacionales (SCN), más los menores de 12 a 17 años en trabajo intensivo (Encuesta Nacional de Hogares - ENAHO y la Encuesta sobre el Trabajo Infantil - ETI,2016).

En el mundo, según la OIT en el 2016 la eliminación del trabajo infantil sigue siendo un desafío considerable. Hay 152 millones de niños, 64 millones de niñas y 88 millones de niños en situación de trabajo infantil; es decir, casi 1 de cada 10 niños en todo el mundo. Poco menos de la mitad de los absolutos realizan trabajos peligrosos que representan un riesgo directo para su salud (OIT, 2017).

A nivel de América Latina y el Caribe según (OIT y CEPAL,2018), “Entre 2012 y 2016, la región mostró una reducción de 17% en la tasa de trabajo infantil y de 35% en la de trabajo infantil peligroso. En otras palabras, dos millones de niños, niñas y adolescentes dejaron de trabajar en nuestros países en ese período”.

Según el Ministerio de Trabajo y Promoción del Empleo (ENAHO y ETI,2016).si bien se ha visto que la tasa del Trabajo Infantil ha ido disminuyendo desde el 2012, aunque las cifras aún siguen siendo elevadas, uno de cada cuatro niños se encuentra ocupado, uno de cada nueve se encuentra en situación de trabajo infantil y uno de cada veintisiete realiza trabajos peligrosos.

Si analizamos los indicadores proporcionados por el INEI(2018) en relación a los objetivos de desarrollo sostenible(ODS), encontramos de que la proporción de niños, niñas y adolescentes de 5 a 17 años en actividad económica ha ido disminuyendo en un 17.7 % desde el 2012 hasta el 2018, aunque si observamos un panorama más actual como desde el 2015 vemos que solo disminuyo un 1.1% lo que indica que aún no se ha logrado atacar este problema en diferentes zonas del país mostrándonos que no se están obteniendo los resultados esperados de los programas que actualmente se están aplicando , como lo que ocurre en el 2018 para este encabezado de departamentos Huancavelica(64.6%), Cajamarca(62.6%), Apurímac(54.3%), Amazonas(47.7%), Pasco(46.6), Áncash(45.9%) y Huánuco(45.8%), este último en el cual nos centraremos puesto que desde el 2012 se encontraba entre los primeros departamento con mayor trabajo infantil, además de que las estimaciones hacia Huánuco en el 2015 según MTPE con respecto a los indicadores dados por el INEI han diferido posiblemente por el uso de diferentes encuestas para cada estimación 65% y 51.9% respectivamente. Debido que no solo Perú sino también todo América Latina y el Caribe se encuentra en el mismo rumbo, la OIT y CEPAL presentaron un informe en el 2018 donde proponen un modelo de identificación del riesgo de trabajo infantil con el fin de poder diseñar políticas preventivas a nivel subnacional, mediante la obtención de estos factores, que en la investigación fue realizada para Brasil se encontró las características que están involucradas con el trabajo infantil con cada estado del país.

1.2. Definición del problema de investigación

Observamos que es de interés proporcionar investigación continua sobre este problema y a su vez de dar un enfoque estadístico con el fin de apoyar con el desarrollo de resultados no oficiales para sugerir o refutar estudios que realicen entidades del estado, Por lo expuesto, en este trabajo abordaremos la caracterización del trabajo infantil en el Perú para el año 2019 utilizando la información de la ENAHO 2019.

1.3. Formulación del problema

Luego de analizar la problemática se plantean la siguiente pregunta de investigación:

1.3.1. Pregunta general

- ¿Cuáles son las características asociadas al trabajo infantil en el Perú para el 2019?

1.3.2. Preguntas específicas

- ¿Cuáles son los factores más importantes para caracterizar al trabajo infantil en el Perú para el 2019?
- ¿Cuáles son los perfiles asociados a los trabajadores infantiles en el Perú para el 2019?

1.4. Objetivos

1.4.1. Objetivo principal

- Caracterizar el trabajo infantil en el Perú para el 2019.

1.4.2. Objetivos específicos

- Identificar los factores más importantes para caracterizar al trabajo infantil en el Perú para el 2019.
- Determinar los perfiles asociados a los trabajadores infantiles en el Perú para el 2019.

1.5. Justificación

1.5.1. Justificación Teórica

La identificación de las características asociadas al trabajo infantil puede utilizarse para el contraste con investigaciones pasadas que sean aplicadas en el mismo contexto, a su vez permitirá tener una visión más actualizada de lo que actualmente se dispone en investigaciones sobre este problema en el Perú, con lo cual se obtiene relaciones de las características de familiares como la educación de estos y el efecto sobre esta problemática, se verá las características tal como la pobreza de la vivienda puede contribuir o afectar a la situación del menor en Perú en el año 2019.

1.5.2. Justificación Práctica

Todo investigador interesado en las características del trabajo infantil comprende la complejidad de la problemática entorno a los menores y los resultados de esta investigación les servirá en la extracción de información relevante para la toma de decisiones, además de ver una perspectiva técnica de las características de los métodos utilizados para la toma de decisiones.

Todo lo comentado ayuda a incentivar al diseño de nuevas medidas de prevención y erradicación del trabajo infantil en el Perú.

2. ESTADO DEL CONOCIMIENTO

Para la mejor comprensión del siguiente estudio se presenta algunas investigaciones relacionadas, las bases teóricas de los temas tratados y los conceptos involucrados:

2.1. Antecedentes

2.1.1. Antecedentes Internacionales

En un informe de la OIT y CEPAL (2018), se implementó una metodología para identificar los factores asociados al trabajo infantil, la cual fue utilizada para una aplicación en Brasil usando los datos extraídos de la encuesta Pesquisa Nacional por Amuestra de Domicilios (PNAD) 2011 como el Censo de Población y Vivienda de 2010 siendo su población objetivo los menores entre 10 a 17 años; con el fin de aplicar el modelo de regresión logística propuesto en este informe. En la aplicación se usó características individuales, asociadas al hogar y relacionadas al jefe y cónyuge del hogar, haciendo uso de las variables sexo, edad, zona, etnia/raza, migración, asistencia educativa, número de personas en el hogar ,tipo de familia, educación de padres, ocupación de padres, tipo de empleo del jefe de hogar, contrato laboral del jefe de hogar y el ingreso del hogar; donde obtuvieron una tasa de acierto de 85% siendo este considerado satisfactorio para su aplicación, además de obtener un McFadden's R^2 de 0.22 el cual fue bien recibido, también se encontró que todas las variables consideradas era significativas a nivel nacional y ellos también aplicaron este modelo sobre los estados individuales viendo que la asistencia a la escuela es un factor protector en cada uno de los estados para que los NNA no se encuentren en riesgo de encontrarse en una situación de trabajo infantil; por otro lado, con la zona rural aumenta la probabilidad del trabajo infantil en los estados, aunque en los estados donde se visualiza mayor número de zonas urbanas no se ven tan afectados.

Como conclusión del informe, se indica que el hacer uso del modelo de regresión logística tiene una simpleza relativa del cual facilita su uso para los países de América Latina y el Caribe, resaltando que hacer uso de los modelos para dar información subnacional como es el caso de los estados no deben hacer uso como cifras oficiales, además de aclarar también las limitaciones de las encuestas y la importancia de la caracterización de los territorios. Con el cual se tomó como referencia para informar mediante el presente estudio, que modelo es el adecuado y si este se adapta al comportamiento peruano.

En el artículo de Khatab, K., Raheem, M., Sartorius y B.,Ismail, M.(2019) su objetivo era identificar los factores sociodemográficos, económicos y geoespaciales asociados con la participación laboral de los niños, además este estudio fue usado con la encuesta de salud y demográfica de Egipto del 2014 en el cual poseían 20560 menores entre 5 a 17 años que se encontraban en alguna actividad económica, dentro o fuera de casa; en este realizaron el análisis con modelos bayesianos multivariados geo aditivos, el trabajo infantil fue discretizado como niños que trabajan menos de 16 horas, entre 16 a 45 horas y mayor a 45 horas, para la cual se usó un modelo multinomial y se hizo uso de técnicas de imputación múltiple mediante cadenas de Márkov y Monte Carlo. Se uso las variables sexo, edad, residencia, tamaño del hogar, índice de riqueza, educación de los padres, supervivencia de padres, castigos psicológicos, físicos y físicos severos. Encontrando que las niñas (en referencia con la categoría de los niños) que trabajaban al menos 16 horas eran más probables de estar en situación de trabajo infantil, que las niñas que trabajan entre 16 a 45 horas y niños de madres sin una educación formal, que no estén en trabajos peligrosos sin considerar las horas de trabajo, son más probables de estar involucradas en trabajo infantil comparado con los que sus madres tenían mejor nivel de educación. Además de que los niños que reportaron agresión tanto física como psicológica también son más probables de estar en esta problemática, finalmente mencionan que Egipto del noroeste tiene mayor probabilidad de que los niños de otra región en estar en situación de trabajo infantil mientras que los niños que viven en Delta son más propensos a trabajar en situaciones peligrosas. Ellos concluyen que la influencia de tanto factores sociodemográficos como económicos son significativos sobre su problemática, además de que sugiere en base a los efectos espaciales dar más atención a las áreas con mayores tasas de trabajo infantil como Nile Delta, Upper Egypt, and Northeastern Egypt. Este trabajo tiene la similitud de haber implementado un método que no se había usado en investigaciones anteriores.

2.1.2.Antecedentes Nacionales

Saenz, C., Lazo, J., López, K. y Bravo, E. (2017) en su artículo el cual tiene como objetivo comparar las técnicas de regresión logística y de redes neuronales para predecir el trabajo infantil en el Perú haciendo uso de la Encuesta Nacional de Hogares del año 2014 sin hacer uso de los tres primeros meses puesto que eran vacaciones y argumentan que retiraría la estacionalidad a su vez no haría intuir que los menores dejaron de estudiar por trabajar, además de solo considerar a los niños entre 12 y 17 años; con lo cual usaron los modelos de regresión logística y redes neuronales haciendo uso de particiones para testear y validar los resultados, hicieron uso de 17 variables obteniendo como resultados que en general la redes neuronales superaba en la capacidad predictiva con respecto al modelo logístico el cual hizo uso de las 17 variables y en el modelo logístico solo uso las variables(9) con significancia y cumplían con los supuestos del modelo, indican que el hacer uso de la otra parte de variables es relevante para la predicción lo que no ocurre en la regresión logística, además adicionan que las redes neuronales captura las relaciones no lineales entre factores; concluyendo que tanto los indicadores geográficos, niveles de ingreso , sexo, composición familiar y nivel de educación son significativos al predecir el trabajo infantil, adicionando que las redes neuronales puede ayudar al gobierno a tomar decisiones eficientes dado que su estudio mostro que provee mayor capacidad predictiva que el modelo logístico. Al hacer esta comparación es similar a mi investigación en el sentido que hace uso de la encuesta ENAHO y está realizando un nuevo método para abordar esta problemática.

Aliaga, L., Guabloche, M., y Villacorta, M. (2009) en su tesis para el grado de magister, presenta como objetivo determinar los factores del trabajo infantil en el Perú para lo cual hizo uso en su investigación de la ENAHO 2007 siendo 18812 menores entre 6 a 17 años , tomando en consideración variables individuales como el sexo, idioma materno, asistencia al colegio, desayuno, nivel educativo, área, niño trabajador; además de variables del hogar como el nivel educativo del jefe del hogar , la unidad familiar de bienes y servicios , el nivel de pobreza y la demanda de trabajo infantil por parte del hogar(Si el padre trabaja en Agricultura); en esta investigación hacen uso del Modelo Logit a todo el Perú y además de estimar por áreas el modelo. Como resultados se tuvo a nivel nacional que las niñas tienen menor probabilidad de trabajar en un 21.5 % con respecto a los niños, además si ~~es que~~ el jefe del hogar no tiene ningún nivel educativo o tiene primaria o secundaria , la probabilidad de que el niño trabaja se incrementa en un 142.4% ,116.3 y 86.3% respectivamente concordando con la literatura presentada; a nivel de costa rural algunos niveles educativos no

ayudan a explicar el trabajo infantil, por otro lado en el modelo de la selva Rural algunos de los idiomas tampoco son relevantes, dando a entender que hay variabilidad de factores por cada área . Como conclusiones se tuvo que si se tiene diversidad geográfica e intercultural los determinantes del trabajo infantil siguen el camino por parte económica, aspectos culturales, pobreza y bajos niveles de instrucción. Al igual que esta tesis también se aplicará en base a la información de la ENAHO, aunque con información actualizada al 2019 y metodología diferente.

2.1.3.Antecedentes Locales

Por otro lado, Benites, S., Pereda, V., Vicuña, J., y Yupari, I. (2013) tienen como objetivo en su artículo el analizar los factores que determinan la situación laboral y su efecto en las condiciones de vida de los niños y adolescentes del Distrito de Víctor Larco Herrera, para lo cual uso una muestra de 1597 niños y adolescentes; el cual tiene un diseño descriptivo, transversal y de tipo no experimental. Y obtuvieron que el 14% realizaba actividades laborales, las características que fueron encontradas como las principales fueron el género, la edad, el número de hermanos y las personas que trabajan en el hogar y concluyeron que las condiciones de vida del niño y el adolescente tienen efectos negativos.

2.2. Bases teóricas

2.2.1.Factores y técnicas usadas

Con el fin de caracterizar el trabajo infantil, la mayoría de investigadores del área de sociología han realizado una ardua investigación sobre los posibles factores que caracterizan a estos individuos e incluso han aplicado diferente clases de modelos para alcanzar este objetivo, en el Perú las investigaciones son similares teniendo como referentes actuales las de MTPE en el 2015 donde aplicaron un modelo de regresión logística para poder encontrar estos factores y otros que también aplicaron este modelo adicionando el modelo probit (Milagros Cayo, 2012), Luz, Maria y Milena (2009), Ray (2000) ,Cortez y Gil (2000) además del modelo logístico aunque no se incluyeron en ninguno la verificación de los supuestos que siguen estos modelos, cabe decir que también son estos los que se han estado aplicando internacionalmente con ciertas variantes como el uso de respuesta multinomial o el uso de diferentes métodos de estimación, siendo el 2014 donde Christian, Juan, Karla y Edgardo donde presentaron su artículo comparando la regresión logística y técnicas de redes

neuronales; otro modelo aplicado fue el de árboles de decisión en el estado de Tocantins en Brasil por Diego, David y Michel (2015).

Tanto OIT y CEPAL (2018), Cayo (2018) ; Aliaga, Guabloche y Villacorta (2009); Grupo de Análisis para el Desarrollo y Consorcio de investigación económica y social (CIES y GRADE, 2011) entre otros (Anexo 3), al encarar su problemática utilizaron el sexo de los menores, la edad de los menores, el área de residencia del hogar de los menores y el nivel educativo de los padres para el ajuste, por otro lado un artículo en el Perú de Saenz, Lazo, López-Yucra y Bravo. (2017) incluye otros factores como parte de su comparativa como razones del número de miembros del hogar, si el menor es cabeza de hogar o también el tipo de la vivienda.

2.2.2. Árbol de decisión

En la literatura encontrada se aprecia que se ha hecho uso de la técnica de regresión logística para la mayoría de los casos y esos otros pocos están asociados al modelo probit, para introducir una nueva manera de caracterizar el trabajo infantil puede ser usando árboles de decisión el cual nos permita ser más flexibles con respecto a los supuestos del modelo como en el caso de la regresión logística; un artículo presentado por Rodrigues, Diego. (2015) dio un impulso para el uso de esta técnica, pero enfocado en las familias.

a. Definición de árboles de decisión

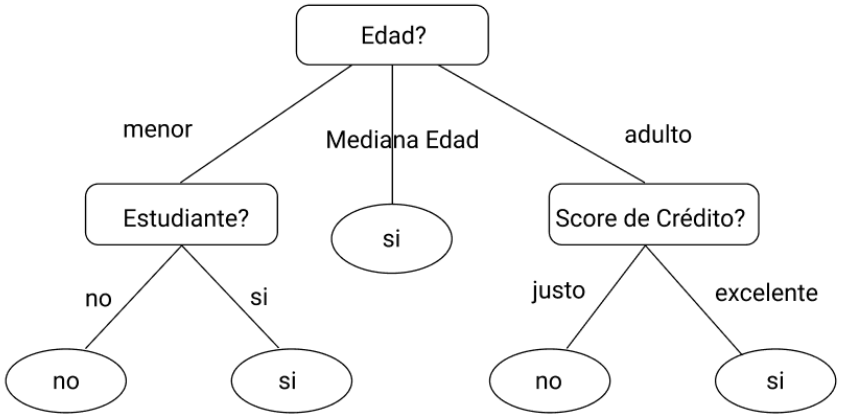
Un árbol de decisión es una secuencia de condiciones expresada como un diagrama de flujo con la estructura de un árbol donde cada nodo interno expresa una evaluación sobre un atributo o característica, cada rama expresa el resultado de esta evaluación, cada hoja del nodo (nodo terminal) resulta en una etiqueta de la clase (decisión final) que deseamos clasificar el cual es entrenado o ajustado a través de la estructura jerárquica del árbol. (Han, Kamber y Pei, 2012). Según Gironés, Casas, Minguillón y Caihuelas (2017) los árboles de decisión son modelos utilizados en minería de datos que subdividen el espacio de las variables explicativas para generar regiones disjuntas, para que así cada elemento pertenezca únicamente a una misma región y si dentro de una región existen elementos de distintas clases estas se subdividen formando regiones más pequeñas hasta particionar todo el conjunto de entrada.

b. Componentes

Un árbol está compuesto de nodos y hojas, los nodos internos son los representados en el siguiente gráfico como los rectángulos, además el elemento superior es llamado la raíz, y las hojas de los nodos los óvalos. Los nodos están asociados a condiciones las cuales crean dos regiones y las hojas terminales contiene los elementos de una sola clase perteneciente a una región pura puesto que ya no existe elementos de diferentes clases.

Figura 1.

Estructura de árbol de decisión



c. Criterios de construcción de un árbol de decisión

Según Gironés, Casas, Minguillón y Caihuelas (2017) muestra cuatro criterios al momento de construir un árbol de decisión

- Criterio de parada, indica cuando se debe dejar de seguir seleccionando nodos para subdividirlos.
- Criterio de selección, se evalúa que nodo va a ser seleccionado para realizar la partición, pero si se particiona como arboles de decisión completos no será relevante este criterio.
- Criterio de clasificación, indica que clase o etiqueta se asignará al nodo hoja, esto es conseguido mediante la minimización del error de clasificación o de regresión.
- Criterio de partición, indica el camino a seguir para poder dividir un nodo en uno o más.

En la teoría presentada en Hastie, T., Tibshirani, R., y Friedman, J. H. (2009) mostraron los correspondientes resúmenes basados en Breiman, L., Friedman, J., Stone, C. J., y Olshen, R. A. (1984) con el algoritmo CART.

El algoritmo CART es definido como un proceso de crecimiento recursivo para respuestas binarias el cual permite tener como variables predictoras tanto cuantitativas como cualitativas (Lewis, R. ,2000). Y el procedimiento resumido por Hastie .et al,(2013) se divide en 4 fases.

A. Mediante el particionamiento binario recursivo se hace crecer el árbol utilizando todas las observaciones, parándolo solo cuando los se cumple el criterio de parada (stopping criteria), como el árbol alcance un número determinado de nodos terminales (tree depth), que el nodo a ser particionado alcance un mínimo de número de observaciones (min split) u otros condiciones son mostrados en (Rokach y Maimon, 2014,p.19).
B. Aplicamos la poda mediante el coste de complejidad al árbol más grande (con mayor profundidad) para obtener una secuencia de los mejores subárboles, como función de α.
C. Se usa validación cruzada con K-particiones para escoger α, escogiendo el que minimice el error promedio.

Se define a x_i como un vector de p variables predictoras, cada una con n observaciones y y_i son los valores de la variable respuesta y:

En el primer paso, el **particionamiento binario recursivo** (A) realiza una división del espacio de las variables explicativas, lo que significa que el conjunto de los posibles valores de $X_1, X_2, \dots X_p$ son distribuidos en M distintas regiones $R_1, R_2, \dots, R_M$

Para cada observación que se encuentre en la región R_m, se realiza la misma predicción, el cual dependerá si la variable dependiente es cuantitativa para lo cual se usaría la media de los valores de las respuestas en esa región o si la variable respuesta es binaria se le asignara la clase mayoritaria.

d. Arboles de Regresión

Si la variable respuesta toma valores reales, se modela basado en una constante a_m en cada región, donde $I(.)$ es la función indicadora

$$f(x) = \sum_{m=1}^{M} a_m I(x \in R_m) \qquad (1)$$

Si se toma como criterio de minimización la suma de cuadrados $\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$, se obtiene que el mejor $\hat{a}_m$ es el promedio de y_i en la región R_m, donde el número de observaciones en la m-ésima región esta denotado como n_m.

$$a_m = \frac{1}{n_m} \sum_{x_i \in R_m} y_i \tag{2}$$

La mejor partición según la suma de cuadrados (Hastie et al., 2009) es computacionalmente factible, por ende, el algoritmo de particionamiento binario recursivo toma en consideración todos los p predictores y todos los posibles valores de corte punto de corte s para cada predictor, entonces se define para cualquier j y s un par de semiplanos.

$$R_1(j,s) = \{X | X_j \leq s\} \; y \; R_2(j,s) = \{X | X_J > s\} \tag{3}$$

Por lo que se busca encontrar la variable j y el punto s que solucione el problema de minimización (4)

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - a_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - a_2)^2 \right] \tag{4}$$

La minimización interna es resuelta para cualquier j y s por

$$\hat{a}_1 = \frac{1}{n_1} \sum_{x_i \in R_1(j,s)} y_i \quad y \quad \hat{a}_2 = \frac{1}{n_2} \sum_{x_i \in R_2(j,s)} y_i \tag{5}$$

Para cada partición de la variable, el cálculo del punto de partición s es muy rápido iterando sobre todas las observaciones y calcular el par (j,s) de la misma manera. Por lo que, habiendo encontrado la mejor partición en el nodo raíz, se particiona los datos en las regiones resultantes y se repite el proceso en cada una de las dos regiones y así para todas las regiones resultantes.

Se define las siguientes expresiones

$$\begin{aligned} n_m &= \#\{x_i \, \epsilon \, R_m\} \\ \hat{a}_m &= \frac{1}{n_m} \sum_{x_i \in R_m} y_i \\ I_m(T) &= \frac{1}{n_m} \sum_{x_i \in R_m} (y_i - \hat{a}_m)^2 \end{aligned} \tag{6}$$

donde $I_m(T)$ es la medida de impureza para los árboles de regresión la cual es la varianza dentro del nodo y n_m es el número de observaciones en la región m.

e. Árboles de Clasificación

Basados en la teoría de Rokach y Maimon (2014), Ma, X. (2018) y en el resumen de Hastie et al. (2009); si la variable respuesta toma valores como $1,2,\ldots,K$, los cambios necesarios comparados con los árboles de regresión estarían relacionados con el criterio de partición de los nodos; el contraste es en cómo se define la medida de impureza que el caso de regresión es mediante $I_m(T)$ en (6) lo cual difiere del problema de clasificación.

En el nodo m, se representa la región R_m con n_m observaciones, entonces

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i = k), \tag{7}$$

es la proporción de la clase k en el nodo m o como la probabilidad que un individuo u observación sea clasificado como clase k en el nodo m. Se clasifican las observaciones en el nodo m a la clase mayoritaria $k(m) = arg\max_k \hat{p}_{mk}$.

Las diferentes medidas de impureza $I_m(T)$ para árboles de clasificación son mostradas en (Breiman et al.,1984) y son las siguientes:

- Entropía Cruzada

$$I_m(T) = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk} \tag{8}$$

También conocida como entropía en Rokach y Maimon (2014), esta mide el grado de desorden de una distribución de elementos en el nodo m y la entropía es cero si todos los elementos son de una misma clase midiendo la impureza de los nodos.

- Coeficiente o Índice de Gini

$$I_m(T) = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{9}$$

Este coeficiente según Rokach y Maimon (2014) mide la divergencia entre la distribución de probabilidades de los atributos de los nodos, es decir el grado de pureza de un nodo con respecto a las otras clases o probabilidad de no sacar dos registros de la misma clase de un mismo nodo. Otra interpretación dada por Hastie et al. (2009) es que, si el valor con el cual se clasifica en cada nodo son de la forma 1 y 0 el coeficiente de Gini para el nodo es la varianza de estos valores, Si los datos, son clasificados correctamente el índice de Gini toma valores cercanos a 0.

- Error de clasificación (Misclassification error)

Tiene una interpretación similar al del coeficiente de Gini, aunque en esta medida se mide la impureza basado en la proporción del nodo con la clase mayoritaria.

$$I_m(T) = \frac{1}{n_m} \sum_{i \in R_m} I(y_i \neq k) = 1 - \hat{p}_{mk} \qquad (10)$$

f. Poda (Pruning)

La profundidad de árbol de decisión será muy grande si no se limita el número de particiones a realizar lo que causaría sobreajuste, y una de las técnicas utilizadas es realizar la partición si es que el decrecimiento de la suma de cuadrados excede cierto límite, pero este si bien es corto podría generar una partición aparentemente mejor aunque en realidad no proporcione valor; por lo tanto la estrategia preferida es parar el particionamiento de los nodos cuando el árbol alcance un mínimo de nodos(tree depth) , por lo que es un criterio de poda.

Sea T_0 un árbol de profundidad alta, un subárbol $T \subset T_0$ que puede ser obtenido de la poda (pruning), indicamos a los nodos terminal como m, con el nodo m representando la región R_m, Además $|T|$ indica el número de nodos terminales en T.

Entonces se define el coste de criterio de complejidad

$$C_\alpha(T) = \sum_{m=1}^{|T|} n_m I_m(T) + \alpha|T|, \alpha \geq 0 \qquad (11)$$

Donde α es definido como el parámetro de costo de complejidad y es un número real mayor o igual que cero. Por lo que el árbol final seria $T_{\hat{\alpha}}$, en el cual para valores grandes de α el árbol es muy pequeño y grande en el sentido opuesto; al considerar un árbol de cada tamaño mediante validación cruzada obtendríamos una aproximación del parámetro de coste de complejidad $\hat{\alpha}$ el cual posee la misma interpretación y se logra minimizar la función (7).

g. La Matriz de Perdida (The Loss Matrix)

Para enfrentar las consecuencias de observaciones incorrectamente clasificadas se define un matriz $\boldsymbol{L} \in \mathbb{R}^{KxK}$, siendo $L_{kk'}$ la perdida obtenida por clasificar como clase k las observaciones de clases k' y no existe pérdida por una correcta clasificación en $L_{kk} = 0$ para todo k.

Para incorporar esta pérdida en el proceso de modelamiento guiados por Therneau, T., Terry, E. (2019) y Hastie et al. (2009) se define una modificación al índice de Gini

$$I_m(T) = \sum_{k \neq k'} L_{kk'} \hat{p}_{mk} \hat{p}_{mk'} \qquad (12)$$

Esta modificación es válida para la clasificación de k>2 clases, en el caso binario solo de pondera por $L_{kk'}$.

h. Validación Cruzada (Cross Validation)

Therneau y Atkinson (2019) explican que la validación cruzada para obtener el mejor parámetro de complejidad se realiza dividiendo el conjunto total de observaciones en k particiones, sugiriendo k= 10 (Hastie et al. (2009)), siendo cada partición $G_1, .., G_i, .., G_k$ de tamaño k/N y en particular para encontrar el óptimo parámetro de complejidad (α), se ajusta el modelo completo sobre la data exceptuando la partición G_i y se determina el riesgo o error para cada costo de complejidad definido en forma de secuencias (0.001,0.01,...,0.1), luego se suma el error por cada parámetro fijado a lo largo de las particiones obtenidas, se elige el valor de α que minimice el promedio del error o error relativo (estandarizado).

Figura 2.

Validación Cruzada Gráficamente

i. Importancia de Variables

La importancia relativa de las variables para el caso de regresión presentada por Breiman et al. (1984) es dada:

$$\mathbb{I}_\ell^2(T) = \sum_{t=1}^{J-1} i_t^2 \, I(v(t) = \ell), \ell = 1, \ldots, p \qquad (13)$$

Donde v(t) es la variable a la cual corresponde el nodo interno t, ℓ son las variables predictoras e i_t^2 es la ganancia máxima obtenida entre los nodos internos definida como el máximo valor de la expresión (4) a minimizar, en general, la importancia de

la variable X_ℓ es la suma de las mejoras(ganancias) al cuadrado a lo largo de todos los nodos internos para los cuales fue escogido para particionar la variable.

Sandri, M., y Zuccolotto, P. (2008) formalizaron lo presentado por diferentes autores para el caso de **clasificación**, el término i_t^2 varia a Δg_t el cual es conocido como la reducción de impureza (o Ganancia de Gini) en el nodo t

$$\Delta g_t = I(t) - I(t_l, t_r) \tag{14}$$

donde t_l *y* t_r son los nodos hijos del nodo t evaluado, y la expresión $I(t_l, t_r)$ *se expresa como:* $I(t_l, t_r) = I(t_l)\, {}^{n_{1\cdot}}/_{n_1 + n_2} + I(t_r)\, {}^{n_{2\cdot}}/_{n_1 + n_2}$

debido a que la importancia de cada variable puede tomar un valor máximo en su rango desconocido, se recurre al escalamiento con un máximo de 100. (Ma, X., 2018) sugiere nombrar a las variables con mayor importancia como "mejores predictores" a comparación de predictores significativos para diferenciar la importancia entre en CART y la importancia bajo métodos tradicionales estadísticos como regresión lineal múltiple.

j. Evaluación del modelo

La calidad o bondad del ajuste que posee un modelo se realiza mediante las predicciones generadas para el conjunto de datos X. (Han, Kamber y Pei, 2012). Una gran variedad de estadísticos para la evaluación de modelos de clasificación es presentada por (Henley, Golden y Kashner, 2020)

Matriz de confusión o Matriz de clasificación

Basado en (Myers, Montgomery y Vining, 2002) y (Zaki y Meira, 2014) se define la matriz de confusión como una tabla de contingencias de las clases observadas (variable respuesta) y la clase predicha.

Tabla 1

Tabla de Clasificación

	Clase verdadera u observada	
Clase predicha	Positivo(P)	Negativo(N)
Positivo(P)	Verdadero Positivo (TP)	Falso Positivo (FP)
Negativo(N)	Falso Negativo (FN)	Verdadero Negativo (TN)

Cada entrada de la matriz de clasificación es llamada con un nombre especial según Henley, Golden y Kashner (2020) como sigue:

- Verdaderos Positivos (TP): El número de puntos que han sido clasificados correctamente como positivas

$$TP = n_{11} = |\{x_i|\widehat{y_i} = y_i = P\}|$$

- Falsos Positivos (FP): El número de puntos son clasificados como positivos, pero en realidad pertenecen a la clase negativa.

$$FP = n_{12} = |\{x_i|\widehat{y_i} = P \ \wedge \ y_i = N\}|$$

- Falsos Negativos (FN): El número de observaciones clasificados como negativos, pero en realidad eran positivos.

$$FN = n_{21} = |\{x_i|\widehat{y_i} = N \ \wedge \ y_i = P\}|$$

- Verdaderos Negativos (TN): El número de observaciones que son clasificados correctamente como negativos.

$$TN = n_{22} = |\{x_i|\widehat{y_i} = y_i = N\}|$$

Métricas derivadas de la matriz de confusión

- Tasa de Error: Proporción de clasificaciones incorrectas

$$Error\ Rate = \frac{FP + FN}{n}$$

- Exactitud: Proporción del total de clasificaciones positivas y negativas correctas.

$$Accuracy = \frac{TP + TN}{n}$$

- Sensibilidad: Proporción de clasificaciones positivas que son predichos como positivos.

$$Sensitivity = TPR = Recall = \frac{TP}{TP + FN} = \frac{TP}{n_1}$$

donde n_1 es el tamaño de la clase positiva.

- Especificidad: Proporción de clasificaciones negativos que son predichos como negativos.

$$Specificity = TNR = \frac{TN}{FP + FN} = \frac{TP}{n_2}$$

donde n_2 es el tamaño de la clase negativa.

Curva ROC

Como la sensibilidad y la especificidad dependen del punto de corte c. Una mejor manera de ver el ajuste de acuerdo con diferentes puntos de corte es mediante el área bajo la curva ROC (Receiver Operating Characteristic). Este gráfico según (Myers, Montgomery y Vining, 2002) muestra la probabilidad de obtener los valores de interés (sensibilidad) contra la probabilidad de no obtenerlos (1-specificidad) para lo cual se definen las funciones:

$$Sens(c) = P[\hat{y} \geq c | y = P]$$

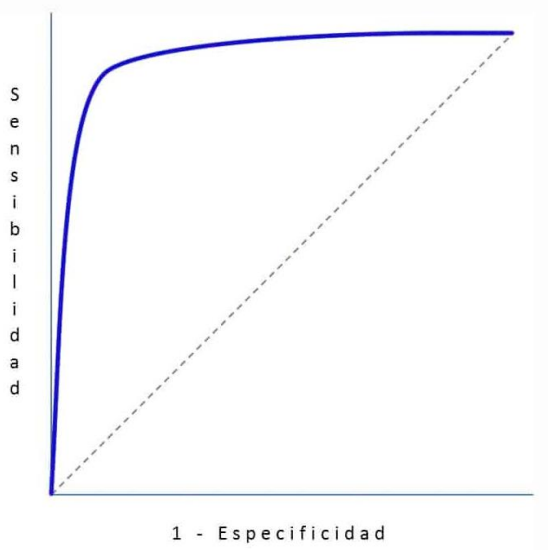
$$1 - Esp(c) = P[\hat{y} \geq c | y = N]$$

Y se obtiene la curva ROC definida como:

$$ROC(.) = \{\big(Sens(c), 1 - Esp(c)\big), c \in (-\infty, \infty)\}$$

Figura 3

Curva ROC



Área bajo la curva ROC (AUC)

Torres, A. (2010) define a AUC como el valor que estima la capacidad de discriminar entre la clase de interés o no interés que tiene el modelo.

$$AUC = \int_0^1 ROC(p)dp$$

Hosmer, Lemeshow, y Sturdivant (2013) sugieren que el área bajo la curva debe seguir las siguientes guías para su interpretación

- $AUC = 0.5$, no discrimina
- $0.7 \leq AUC \leq 0.80$, discrimina muy bien dando el mejor modelo.

k. Ventajas y Desventajas

Según Han, Kamber y Pei (2012), James, Witten, Hastie y Tibshirani(2013) presentan un resumen con ventajas y desventajas al utilizar los árboles de decisión.

Ventajas

- La construcción es sencilla, aunque puede ser costosa computacionalmente con datos de gran dimensión
- Los resultados son interpretables y se puede evaluar la importancia de cada variable.
- Los árboles pueden ser mostrados gráficamente y son fácilmente explicados incluso por alguien que no es experto.
- Se puede combinar variables cuantitativas como cualitativas, siendo invariantes a alguna transformación.
- Se pueden trabajar con valores perdidos utilizando condiciones alternativas.
- La implementación práctica se basa en condicionales del tipo if then else.
- Desafortunadamente, los árboles de decisión no se encuentran al mismo nivel predictivo que los modelos de clasificación más complejos. Sin embargo, mediante la agregación de múltiples árboles de decisión como lo es las técnicas de bagging, boosting y stacking los cuales aumentan la capacidad predictiva del modelo.

Desventajas

- Puede ocurrir que la hoja este particionada de forma desequilibrada, generando nuevas hojas, con una restricción de para, la participación de la hoja es una posible solución.
- Se puede repetir variables en nodo internos, lo cual ocurre debido a la relación no lineal con la variable respuesta; sin embargo, se podría transformar la variable para evitar esos casos.
- Si existe colinealidad o multicolinealidad podrían indicar que la estructura interna es imposible de capturar, por lo que se recomienda capturar otra variable que pueda liberar de este problema.

- Los datos atípicos podrían hacer que el árbol genere ramas profundas, por eso es mejor hacer un tratamiento adecuado a los atípicos.

2.3. Definición de términos

Caracterización: El determinar los atributos, rasgos de alguien o algo de manera que se distinga de los demás.

Factores: Variables que se usan en el estudio para conocer cómo afectan o influyen a la característica de interés o variable respuesta

Perfil: Conjunto de rasgos peculiares que caracterizan a alguien o algo.

Entropía: Medida de incertidumbre en la teoría de la información

Jefe(a) de Hogar: Núcleo familiar en cual provee de información general del hogar y el encargado de brindar información de los menores de 12 años.

3. MÉTODO

3.1. Tipo y Diseño de investigación

El diseño de investigación es no experimental y transversal puesto que la encuesta fue realizada en el 2019; de tipo descriptiva y correlacional según (Bernal,2011), descriptiva dado que se representa los aspectos más característicos de los individuos y correlacional puesto que en la caracterización toma en cuenta las relaciones entre las variables y se analizará las asociaciones entre variables en relación con el trabajo infantil.

3.2. Fuente de información

La información para analizar proviene de una fuente secundaria dado que los datos han sido obtenidos de la ENAHO 2019, en la cual la muestra de la ENAHO es de tipo probabilística, de áreas, estratificada, multietápica e independiente en cada departamento de estudio siendo el tamaño de muestra 36 994 viviendas particulares. En el cual se tuvo como unidad de investigación los integrantes del hogar familiar y con respecto a los menores de 12 años la información la proporciona una persona responsable del hogar). También se excluye a las personas que residen en viviendas colectivas (hoteles, hospitales, asilos y claustros religiosos, cárceles, etc.).

3.3. Variables de estudio

Luego de haber realizado una exhaustiva investigación sobre variables comúnmente utilizados en la realización de trabajos similares (Anexo 3) y obteniendo las que más se asemejan de la ENAHO -2019, presentamos las variables a considerar para este proyecto de investigación.

Tabla 2

Listado de variables de investigación

Variables	Tipo-Escala	Descripción
Situación Laboral del Menor	Cualitativa (Binaria)	Estuvo trabajando la semana pasada
Sexo del Menor	Cualitativa (Binaria)	Sexo del menor
Edad del Menor	Cuantitativa (Razón)	Edad en años del menor
Nro. Miembros del Hogar	Cuantitativa (Razón)	Número de miembros en el hogar
Nro. Menores de Edad	Cuantitativa (Razón)	Menores de edad (< 18 años)
Nro. Mayores de edad en el Hogar	Cuantitativa (Razón)	Menores de edad (>=18 años)
Ultimo Nivel Aprobado por el Menor	Cualitativa (Nominal)	Último nivel aprobado por el menor
Ultimo Nivel Aprobado por el jefe del Hogar	Cualitativa (Nominal)	Último nivel aprobado por el jefe(a) del hogar
Trabajo del jefe del Hogar	Cualitativa (Nominal)	Clasificación Industrial Internacional Uniforme (CIIU) del Padre
Ingreso y Gasto Bruto del Hogar Anual	Cuantitativa-Razón	Ingreso y Gasto anual bruto del hogar medido en soles
Estrato Socio Económico	Cualitativa (Nominal)	Estrato Socio Económico dado por INEI
Situación de Pobreza	Cualitativa (Nominal)	Nivel de pobreza clasificados: Pobre extremo Pobre no extremo No Pobre
Dominio	Cualitativa (Nominal)	Dominio Geográfico: Como Costa Norte, Costa Centro, Costa Sur, Sierra Norte, Sierra Centro Sierra Sur, Selva y Lima Metropolitana
Regiones	Cualitativa (Nominal)	Regiones del Perú

Fuente: Elaboración propia

3.4. Recopilación y organización de datos

La base de datos para esta investigación se extrajo mediante la unión de los módulos (Características del Hogar y Vivienda, Educación, Economía) de la ENAHO 2019 con previa validación de la información. Posteriormente fueron seleccionados los menores de 5 a 17 años en base al objetivo de investigación, los cuales fueron 30857 menores, luego de haber obtenido solo los que poseían información completa se encontró 29962 menores (97%) con registros completos de 6,872,105 como población.

Luego se realizó el análisis descriptivo de los datos obteniendo medidas resumen de las variables a usar, luego se aplicó pruebas de asociación mediante los lenguajes de programación R versión 4.03 y las principales librerías usadas fueron rpart, tidyverse, tidymodels, broom, vip, haven, janitor y funModeling. Asimismo, se realizó la exploración de los resultados del modelo de árbol de clasificación para luego realizar el ajuste del modelo seleccionado para identificar las características asociadas al riesgo de trabajo infantil.

Para el análisis descriptivo se recategorizaron algunas variables como la edad del menor, la cual fue categorizada en Niños (3 - 11 años) y Adolescente (12 - 17) según las etapas de vida presentadas por MINSA, el número de miembros del hogar mayores y menores de edad fueron también categorizados; el dominio geográfico fue reducido a cuatro categorías (Costa, Sierra, Selva y Lima Metropolitana);las categorías del último nivel de educación tanto del jefe(a) de hogar y del menor fueron recategorizadas a una jerarquía superior.

4. RESULTADOS

4.1. Análisis descriptivo

Luego de haber consolidado la información para el análisis se realizó un análisis univariado. En la muestra estudiada (ver Tabla 2), hay un mayor porcentaje de niños comparado al de los adolescentes 52.3% y 47.7%, luego existe mayor presencia de menores con Primaria incompleta (40.6%) como último nivel aprobado del menor seguido de los menores con Secundaria Incompleta (29.9%) y con menor presencia son los menores con Secundaria completa (5%). Y con respecto a los jefes de hogares de los menores, existe un 46.4% de menores con jefes de hogar realizando actividades de Agricultura, ganadería, silvicultura y Pesca, además un 41.2% de los menores provienen de jefes(as) de hogar con Secundaria y 19.9% provienen de hogares con jefes(as) con estudios Superiores; ahora bien, existe un 70.4% de menores que provienen de hogares clasificados como No Pobres mientras que un 5.5% y 24.1% de menores provienen de hogares categorizados como pobres extremos y pobres no extremos respectivamente. Igualmente, los menores que provienen de hogares provenientes de Sierra son un 37.7%, mientras que los menores que se localizan en Lima Metropolitana son un 8.4%.

Por último, con lo que respecta a los miembros del hogar, hay un 61.2% de menores que se encuentran en hogares con 1 a 2 miembros mayores de edad en el hogar y un 16.2% que se localizan en hogares con 3 a 10 mayores de edad miembros.

Tabla 3

Menores de 5 a 17 años según características sociodemográficas. ENAHO 2019.

Características sociodemográficas	**n**	**%**
Sexo del Menor		
Hombre	14100	50.9
Mujer	13601	49.1
Etapas de Vida(Edad)		
Adolescente	13201	47.7
Niño	14500	52.3
Último nivel aprobado del menor		
Primaria completa	2366	8.5
Primaria incompleta	11256	40.6
Secundaria completa	1391	5.0
Secundaria incompleta	8271	29.9
Otros	4417	15.9
Actividad laboral del jefe(a) de hogar		
Agr,gan,sivi y pes	12861	46.4
Otros	14840	53.6

Características sociodemográficas	n	%
Último nivel aprobado del jefe(a) de hogar		
Primaria	9865	35.6
Secundaria	11419	41.2
Superior	5499	19.9
Otros	918	3.3
Pobreza en el Hogar		
Pobre Extremo	1526	5.5
Pobre No Extremo	6685	24.1
No Pobre	19490	70.4
Dominio Geográfico		
Costa	7271	26.2
Sierra	10448	37.7
Selva	7661	27.7
Lima Metropolitana	2321	8.4
Nro. de mayores de edad en el hogar		
[1,2]	16952	61.2
(2,3]	6234	22.5
(3,10]	4515	16.3
Nro. de menores de edad en el hogar		
[1,2]	14311	51.7
(2,3]	6765	24.4
(3,11]	6625	23.9
Nro. de miembros del hogar		
[2,4]	10306	37.2
(4,6]	11126	40.2
(6,20]	6269	22.6

Con respecto al trabajo infantil (Tabla 4), El 13.1% de menores hombres de 5 a17 años trabajan a diferencia de las mujeres que alcanzan 11.4%, en niños en edades de 3 a 11 años el 19,3% se encuentran trabajando y adolescentes solo 4.3%. Con respecto al último nivel aprobado por el menor, primaria completa es el que posee mayor presencia de trabajo infantil con un 21.9% del total en ese nivel, el nivel otros conformado por los niveles educación Inicial, sin nivel y especial tienen una presencia de 8.8%, siendo menor la presencia en el nivel de secundaria incompleta.

En relación con los jefes(as) del hogar, se tiene que del total de menores provienen de jefes(as) que laboran en la agricultura, silvicultura o pesca, el 20.4% trabajan y es la actividad con mayor presencia de trabajo infantil en comparación a otras que tienen solo un 5.1%, asimismo de los menores que provienen de hogares con jefes(as) con primaria un 18.2% trabajan siendo el porcentaje el mayor con respecto a los menores que vienen de hogares con jefes(as) con educación secundaria y superior, sin embargo los nivel con mayor presencia de trabajo infantil se encuentran en los menores de hogares con jefes de hogar con educación inicial y sin nivel los cuales fueron categorizado como otros(19.7%).

De los menores provenientes de hogares clasificados como pobres extremos, pobre no extremo o pobre, el que tienen mayor presencia de trabajo infantil son los menores provenientes de hogares pobres extremos (21%) y un 9% en los que provienen de hogares no pobres. Además, las regiones con mayor presencia de trabajo infantil en los registros se encuentran entre Cajamarca (28%), Huancavelica (26%), Apurímac (24%) y Huánuco (23%), mientras los que poseen menor presencia son Lima (1.4%), Ica (1.3%) y el Callao (0.6%) ver en Anexo 5 - Tabla 6; luego de categorizar el Dominio Geográfico sé observa que de la Sierra es donde hay mayor presencia de trabajo infantil con 19.6% y es en Lima Metropolitana donde hay solo un 1.1%. Finalmente, con respecto a los miembros del hogar, la mayor presencia de trabajo infantil es en la población de menores de hogares con 1 o 2 mayores de edad y la menor presencia es en lo menores de hogares con mayor de 3 mayores de edad en el hogar; el mayor porcentaje de menores que laboran se encuentran en menores que provienen de hogares con más de 3 miembros menores de edad.

Durante la exploración de asociaciones de las variables independientes con el trabajo infantil mediante la prueba Chi-cuadrado se encontró que existe dependencia estadística con todas las variables (Tabla 4) aportando a la decisión de incluir estas variables en el proceso de modelado estadístico.

Tabla 4

Trabajo Infantil en niños de 5 a 17 años según características sociodemográficas, ENAHO 2019.

Características sociodemográficas	Trabaja		No Trabaja			Chi Cuadrado p-valor
	n	%	n	%	Casos	
Sexo del Menor						
Hombre	1847	13.1	12253	86.9	14100	<0.001
Mujer	1546	11.4	12055	88.6	13601	
Etapas de Vida (Edad)						
Adolescente	565	4.3	12636	95.7	13201	<0.001
Niño	2828	19.5	11672	80.5	14500	
Último nivel aprobado del menor						
Primaria completa	132	5.6	2234	94.4	2366	
Primaria incompleta	2463	21.9	8793	78.1	11256	
Secundaria completa	113	8.1	1278	91.9	1391	<0.001
Secundaria incompleta	297	3.6	7974	96.4	8271	
Otros	388	8.8	4029	91.2	4417	
Actividad laboral del jefe(a) de hogar						
Agr,gan,sivi y pes	2630	20.4	10231	79.6	12861	<0.001
Otros	763	5.1	14077	94.9	14840	

Características sociodemográficas	Trabaja		No Trabaja			Chi Cuadrado
	n	%	n	%	Casos	p-valor
Último nivel aprobado del jefe(a) de hogar						
Primaria	1794	18.2	8071	81.8	9865	<0.001
Secundaria	1227	10.7	10192	89.3	11419	
Superior	191	3.5	5308	96.5	5499	
Otros	181	19.7	737	80.3	918	
Pobreza en el Hogar						
Pobre Extremo	321	21.0	1205	79.0	1526	<0.001
Pobre No Extremo	1190	17.8	5495	82.2	6685	
No Pobre	1882	9.7	17608	90.3	19490	
Dominio Geográfico						
Costa	348	4.8	6923	95.2	7271	<0.001
Sierra	2047	19.6	8401	80.4	10448	
Selva	973	12.7	6688	87.3	7661	
Lima Metropolitana	25	1.1	2296	98.9	2321	
Nro. de mayores de edad en el hogar						
[1,2]	2305	13.6	14647	86.4	16952	<0.001
(2,3]	717	11.5	5517	88.5	6234	
(3,10]	371	8.2	4144	91.8	4515	
Nro. de menores de edad en el hogar						
[1,2]	1499	10.5	12812	89.5	14311	<0.001
(2,3]	853	12.6	5912	87.4	6765	
(3,11]	1041	15.7	5584	84.3	6625	
Nro. de miembros del hogar						
[2,4]	1167	11.3	9139	88.7	10306	<0.001
(4,6]	1382	12.4	9744	87.6	11126	
(6,20]	844	13.5	5425	86.5	6269	

4.2. Modelo árbol de clasificación

El método de árbol de clasificación (CART), se aplicó con la finalidad de caracterizar al trabajo infantil en el Perú para el 2019 en base a características sociodemográficas como el dominio geográfico del hogar, la región, la pobreza del hogar, actividad del jefe(a) del hogar, nivel de educación del jefe(a) y del menor del hogar, sexo y edad del menor, miembros mayores y menores de edad en el hogar y el logaritmo natural de los gastos brutos totales anuales del hogar siendo todas significativas con respecto al trabajo infantil mediante la prueba chi cuadrada ya mencionada.

4.2.1.Ajuste del Modelo

Se realizó el ajuste tomando los parámetros por defectos del software los cuales definían un criterio del número mínimo de observaciones en el nodo para no seguir particionando para este caso 20, con una máxima profundidad de 30. Para evaluar el ajuste del modelo se determinó el parámetro de complejidad $\alpha = 0.003$, debido a la disminución del error relativo (Figura 4) sin aumentar el tamaño del árbol evitando el sobreajuste y la complejidad y resultados se presentan en la figura 7.

4.2.2. Evaluación del Modelo

Para la evaluación del modelo se utilizó la curva ROC (Ilustración 4) para representar la sensibilidad debido a la especificidad en base al modelo obtenido, se observa en la que la curva se encuentra achatada hacia la izquierda, debido a la poca proporción de la clase de interés (12%), es por lo que se evaluaran las métricas del modelo con un punto de corte de 0.1 obteniendo los siguientes resultados:

Figura 4

Validación Cruzada, mediante la distribución de parámetros de complejidad y error relativo

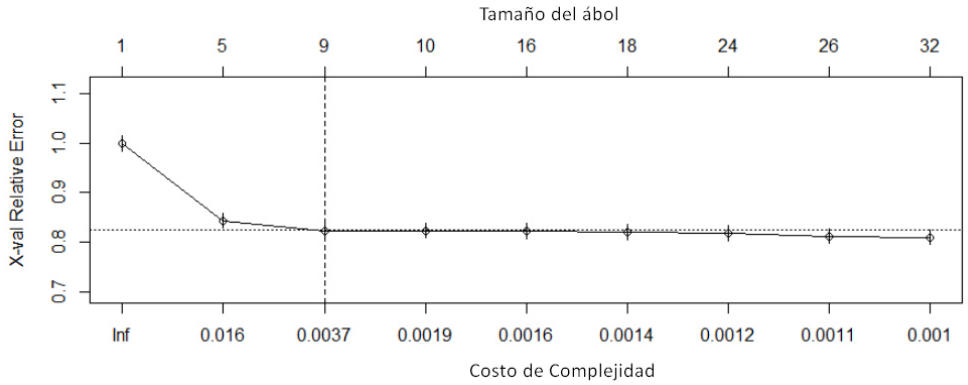


Figura 5

Curva Característica de Operación - ROC

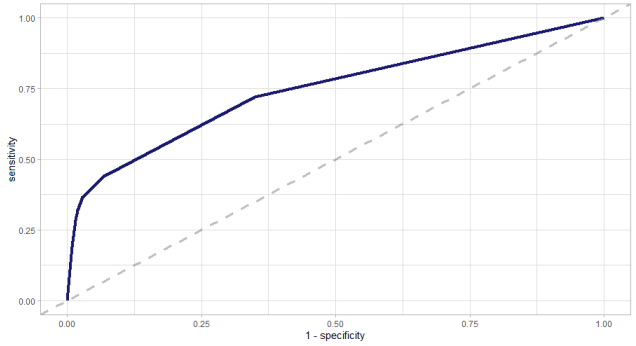


Tabla 5

Matriz de menores según trabajo infantil y predicción basado en árbol de clasificación

Predicción	Trabajo Infantil	
	Si	No
Si	2448	8505
No	945	15 803

Fuente: Elaboración Propia

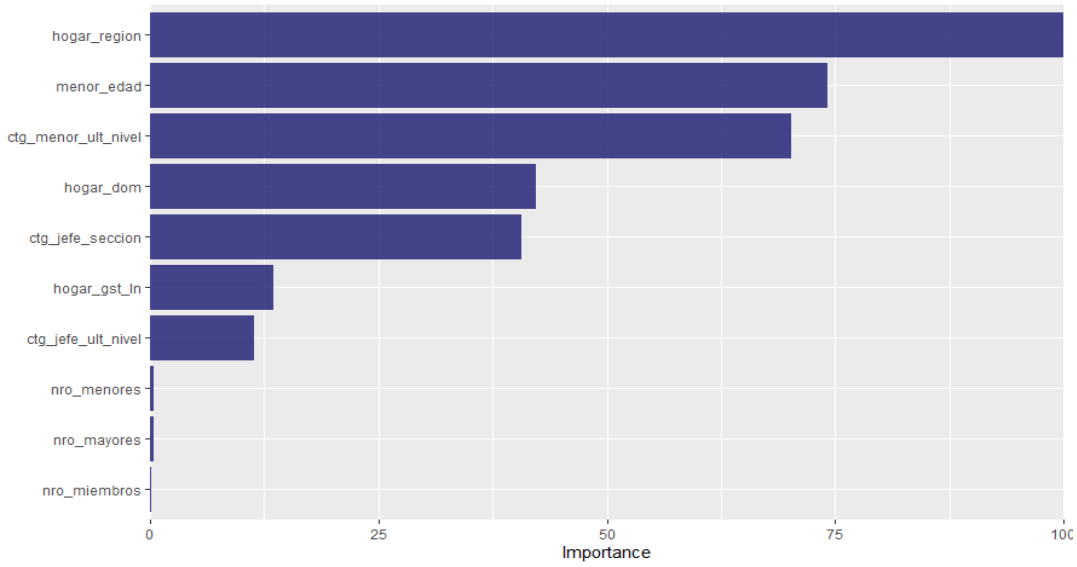
Donde se observa en la Tabla 5 que los falsos negativos obtenidos es de 945 menores, y falsos positivos de 8505, en comparación de los verdaderos positivos y negativos son menores; en lo que respecta a las medidas de evaluación del modelo al momento de caracterizar, se obtuvo una Sensibilidad de 0.72 el cual expresa la proporción de menores en situación de trabajo infantil que fueron clasificadas correctamente como trabajadores, el cual nos da un buen indicador para explicar el objetivo de caracterización del trabajo infantil; sin perder el aspecto discriminante del modelo obteniendo un AUC, exactitud, especificidad y razón de error de 0.75, 0.66, 0.65, 0.79 respectivamente.

4.3. Análisis e interpretación de resultados

A continuación, se presenta la importancia de los factores utilizados en el ajuste del modelo, en base a la participación en los nodos del árbol final.

Figura 6.

Importancia en la caracterización del trabajo infantil



Luego de haber obtenido el árbol final obtenemos las variables explicativas denominadas mejor predictoras, las 25 regiones del Perú para el año 2019 en conjunto son las más importantes para explicar el trabajo infantil en el Perú, siendo las tres regiones con mayor presencia de trabajadores infantiles Cajamarca (28%), Huancavelica (26%), Apurímac (24%). Asimismo, el dominio geográfico del hogar es el cuarto factor que mejor discrimina a los trabajadores infantiles, donde el dominio agrupado de la Sierra es la que mayor presencia de trabajo infantil posee.

La edad del menor se encontró como el segundo más importante, de esta manera se confirma que en 2019 este factor sigue tomando un papel importe en la eliminación del trabajo infantil con respecto a los antecedentes encontrados.

El árbol de decisión obtenido concluye que el último nivel aprobado del menor es el tercer predictor más importante, así pues, se encontró que la Primaria incompleta es el nivel con mayor presencia de trabajo infantil en la muestra y haciendo el papel de la asistencia a la escuela, el cuál es utilizado en previos trabajos.

Respecto a la actividad económica del jefe(a) de hogar, se obtuvo que la importancia es similar al del dominio geográfico, con esto se verifica el hecho de que el menor provenga de hogares con jefes(as) que realicen actividades de Agricultura, silvicultura y pesca sigue aportando al momento de caracterizar al trabajo infantil.

De manera similar al gasto y el ultimo nivel aprobado del jefe(a) de hogar poseen similar importancia, aunque baja a comparación de los antes mencionados y con respecto al número de miembros de los hogares, ya sean menores o mayores, el sexo, además del nivel de pobreza del hogar no tienen suficiente importancia en el Perú para el año 2019 para discriminar a los menores según su situación.

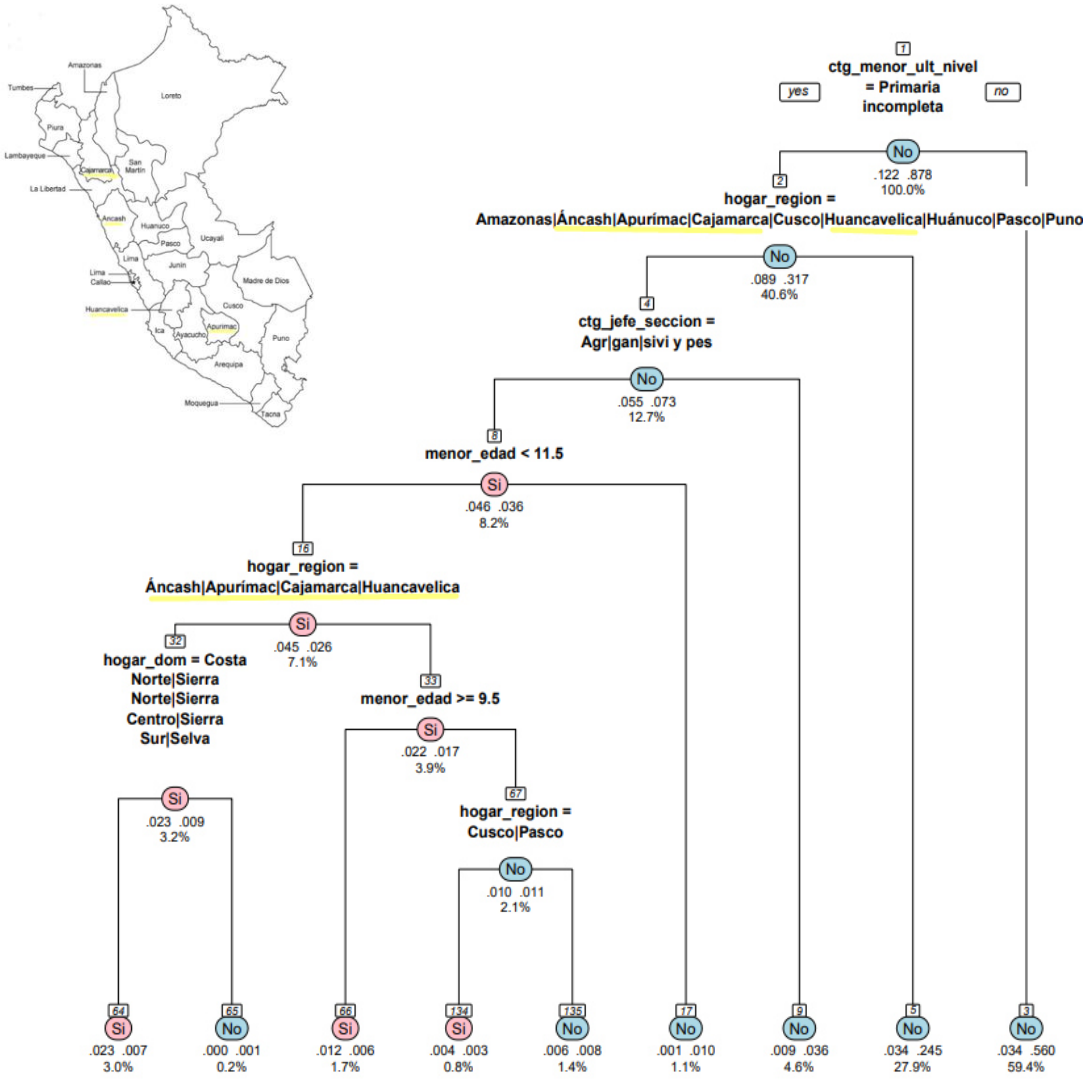
El árbol de clasificación (ver Figura 7) permite caracterizar a los menores en trabajo infantil de la siguiente manera:

Los menores que se encuentran realizando trabajos son aquellos infantes que en su mayoría alcanzaron la primaria incompleta, se ubican en hogares localizados en las regiones de Áncash, Apurímac, Cajamarca y Huancavelica; donde el jefe(a) de hogar labora en actividades de agricultura, silvicultura y pesca, de los menores mencionados un grupo se ubican en la costa norte, sierra y selva, además de ser menores a 12 años y también está el grupo de menores entre 10 a 11 años. De la misma manera, se encontró menores de 9 años que se encuentran realizando trabajos que han alcanzado el nivel de estudios primaria incompleta y viven en las regiones de Cusco y Pasco viven en hogares cuyo jefe de hogar realiza actividad agricultura, silvicultura y pesca.

En consecuencia, también se obtuvieron perfiles asociados a los menores que no trabajan, estos se presentan en el Anexo 5

Figura 7

Árbol de clasificación de menores



5. DISCUSIÓN Y CONCLUSIONES

5.1. Discusión

La situación actual del trabajo infantil en el Perú se ve amenazado dado al contexto actual y a la relación positiva que tienen con la pobreza, esto repercutirá sobre la salud infantil, la economía y las familias, esto es lo que en el estudio se desea caracterizar.

Según OIT, CEPAL (2018) en la aplicación que realizaron en Brasil se tuvo para el 2010 un 12.4% de trabajo infantil entre 10 a 17 años, en relación con Khatab, et al. (2019) en 6.7% en la edad de 5 a 14 años trabajaba, Cayo, M. (2018) para la encuesta de ETI 2015 mostro un 26% de trabajadores infantiles y según Aliaga, et al. (2009) de la muestra de la ENAHO 2007 hubo un 30% de menores en situación de trabajo infantil, mientras que en el presente estudio la proporción es de 12% para la ENAHO 2019 similar al porcentaje encontrado en Brasil en el 2010 e inferior a los porcentajes a nivel nacional del 2007 y 2015.

En ese mismo contexto, la edad, ya sea el ingreso o el gasto del hogar, el nivel educativo del jefe de hogar en Cayo, M. (2018), Saenz, et al. (2017), Aliaga, et al. (2009) resultaron ser significativas al explicar el trabajo infantil, lo cual concuerdan con la investigación. Por otro lado, el número de menores y mayores de edad en Saenz y Cayo encontraron que eran significativas, mientras que en el estudio actual fueron las características que proporcionaron la menor importancia; en adición Saenz,et al.(2017) muestra que el uso de todas las variables disponibles para su estudio ayuda a mejorar la capacidad predictiva de su modelo de redes neuronales, aunque esto no se ve evidenciado en nuestra investigación con árboles de clasificación, esto pudo ser debido a la disminución de porcentaje de la característica de interés en la muestra del 2019 o al uso de mayor observaciones. En comparación a los anteriores autores, se incluyó el último nivel aprobado por el menor para evitar disminuir nuestras observaciones como se hizo en Saenz, et al.(2017), asimismo se usó un modelo no paramétrico y a su vez interpretable para poder categorizar por perfiles a los trabajadores infantiles.

5.2. Conclusiones

La caracterización del trabajo infantil en el Perú fue posible utilizando el árbol de clasificación que permitió concluir:

- Como variables más importantes a las regiones del Perú como la mejor predictora, la edad y el último nivel educativo alcanzado del menor, el dominio geográfico, si la actividad del jefe es relacionada a la agricultura, silvicultura y pesca; por otro lado, con menos nivel de importancia se obtuvo al gasto anual del hogar y el nivel educativo alcanzado por el jefe(a) de hogar, asimismo el número de miembros en el hogar sean menores de edad o mayores y el sexo del menor no son importantes.
- Los perfiles identificados de los menores en situación de trabajo infantil fueron:

  Los menores que se encuentran realizando trabajos alcanzaron la primaria incompleta, y viven en hogares cuyo jefe de hogar realiza actividades de agricultura, silvicultura y pesca, localizados en las regiones de Áncash, Apurímac, Cajamarca y Huancavelica; de los mencionados un grupo se ubican en la costa norte, sierra y selva, además de ser menores a 12 años y también está el grupo de menores entre 10 a 11 años.

  Los menores de 9 años que se encuentran realizando trabajos han alcanzado primaria incompleta y viven en las regiones de Cusco y Pasco viven en hogares cuyo jefe de hogar realiza actividad agricultura, silvicultura y pesca.

- Por último, se encontró que el modelo de árboles de decisión fue eficiente al caracterizar el trabajo infantil encontrando así tres perfiles de los menores en esta situación y otros seis para los que no se encuentran trabajando con una sensibilidad de 0.72 sin incurrir en un modelo complejo o restringido por supuestos.

6. REFERENCIAS BIBLIOGRÁFICAS

Ali, M., Hickman, P., y Clementson, A. (1975). The Application of Automatic Interaction Detection (AID) in Operational Research. Operational Research Quarterly (1970-1977), 26(2), 243-252. https://doi.org/10.2307/3008458

Aliaga, L., Guabloche, M., y Villacorta, M. (2009). Los determinantes del trabajo infantil rural en el Perú y su incidencia sobre la formación del capital humano: Bases para propuestas políticas. http://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/1363

Benites, S., Pereda, V., Vicuña, J., y Yupari, I. (2013). Perú., Factores que determinan la situación laboral y su efecto en las condiciones de vida de los niños y adolescentes del Distrito de Víctor Larco Herrera. Trujillo. UCV-Scientia, 91-104. https://dialnet.unirioja.es/servlet/articulo?codigo=6181520

Bernal Torres, C. A. (2010). *Metodología de la Investigación* (3ra ed.). Pearson.

Breiman, L., Friedman, J., Stone, C. J., y Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cajas, A., Echevarría, J. y Leandro, L (2016). El Trabajo infantil de los niños de la calle: Factores socioeconómicos en la ciudad de Huánuco 2015. Universidad Nacional Hermilio Valdizán. http://repositorio.unheval.edu.pe/handle/UNHEVAL/800

Cayo, M. (2018). Determinantes del trabajo infantil en niños que asisten a la escuela. Pontificia Universidad Católica del Perú. http://tesis.pucp.edu.pe/repositorio/bitstream/handle/20.500.12404/13443/CAYO_VELASQUEZ_MILAGROS_ESCUELA.pdf?sequence=1&isAllowed=y

Comisión Económica para América Latina y el Caribe (CEPAL). (2009) Trabajo infantil en América Latina y el Caribe: su cara invisible. https://repositorio.cepal.org/bitstream/handle/11362/35995/1/Boletin-desafios8-CEPAL-UNICEF_es.pdf

Consorcio de investigación económica y social (CIES) y Grupo de Análisis para el Desarrollo (2011). Trabajo adolescente y deserción escolar en el Perú. http://www.cies.org.pe/sites/default/files/investigaciones/trabajo-adolescente-y-desercion-escolar-en-el-peru.pdf

Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). Journal of the Royal Statistical Society. Series B, 215-242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

ENAHO (2020). Condiciones de vida de la población en riesgo frente a la pandemia del COVID-19. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1745/libro.pdf

ENAHO y ETI (2016). Magnitud y características del trabajo infantil en Perú: Informe de 2015 - Análisis de la Encuesta Nacional de Hogares (ENAHO) y de la Encuesta sobre Trabajo Infantil (ETI). http://white.lim.ilo.org/ipec/documentos/informeti_2015_peru.pdf

Fernando Berzal. *Redes Neuronales & Deep Learning (1era ed.)*. https://deep-learning.ikor.org

Gillo, M., y Shelly, M. (1974) Predictive Modeling of Multivariable and Multivariate Data. Journal of the American Statistical Association 69:347, pages 646-653. https://doi.org/10.2307/2285995

Gironés, J., Casas, J., Minguillón, J., y Caihuelas, R. (2017). *Minería de datos: modelos y algoritmos*. UOC.

Goodfellow, I., Bengio, Y. y Courville, A(2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Han, J., Kamber, M., y Pei, J. (2012). *Data Mining Concepts and Techniques (3era ed.)*. ELSEVIER.

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.

Henley, S. , Golden, R. y Kashner, T. (2020). Statistical modeling methods: challenges and strategies. Biostatistics & Epidemiology, 105-139. https://doi.org/10.1080/24709360.2019.1618653

Hilbe, J. (2009). *Logistic Regression Models (1era ed.)*. CRC Press.

Hosmer, D., Lemeshow, S., y Sturdivant, R. (2013*). Applied logistic Regression (1era ed.)*. Jhon Wiley & Son

Instituto Nacional de Estadística e Informática (2018). PERÚ: SISTEMA DE MONITOREO Y SEGUIMIENTO DE LOS INDICADORES DE LOS OBJETIVOS DE DESARROLLO SOSTENIBLE. Recuperado de

http://ods.inei.gob.pe/ods/objetivos-de-desarrollo-sostenible/trabajo-decente-y-crecimiento-economico.

James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R (1era ed.)*. Springer.

Jared Dean. 2014. *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners (1ra. ed.)*. Wiley Publishing.

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(2), 119-127. https://doi.org/10.2307/2986296

Khatab, K., Raheem, MA., Sartorius y B.,Ismail, M. (2019). Prevalence and risk factors for child labour and violence against children in Egypt using Bayesian geospatial modelling with multiple imputation. PLoS ONE 14(5): e0212715. https://doi.org/10.1371/journal.pone.0212715

Kuhn, M. y Johnson, K. (2013). *Applied Predictive Modeling (1st. ed.)*. Springer.

Ma, X. (2018). *Using classification and regression trees: A practical primer (1ra. ed.)*. Information Age Publishing.

Montgomery, D. C., Peck, E. A., y Vining., G. G. (2012). *Introduction to Linear Regression Analysis (1ra. ed.)*. Wiley.

Muñoz C. (2011). *Cómo elaborar y asesorar una investigación de tesis (1ra. ed.)*. Pearson.

Myers, R., Montgomery, D. y Vining, G. (2002). Myers, R. H., Montgomery, D. y Vining, G. (2002*). Generalized linear models: With applications in engineering and the sciences (1ra. ed.)*. Wiley.

Organismo Internacional del Trabajo (2017). Estimaciones mundiales sobre el trabajo infantil: Resultados y tendencias 2012-2016. https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/documents/publication/wcms_651815.pdf

Organismo Internacional del Trabajo (2018). Trabajo Infantil. Organización Internacional del Trabajo. https://www.ilo.org/global/standards/subjects-covered-by-international-labour-standards/child-labour/lang--es/index.htm

Organismo Internacional del Trabajo y CEPAL (2018). Modelo de Identificación del Riesgo de Trabajo Infantil: Metodología para diseñar estrategias preventivas a nivel local. https://www.iniciativa2025alc.org/sites/default/files/modelo-de-

identificacion-del-riesgo-de-trabajo-infantil_IR.pdf?fbclid=IwAR09IX24aA28RW-ZH-DLirhtLH9_dbR2VLuIbLCZ3Cr-zHP3Bqy-TCXiZT0

Rodrigues, Diego. (2015). Exploring Social Data to Understand Child Labor. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.7763%2FIJSSH.2015.V5.416

Saenz, C., Lazo, J., López, K. y Bravo, E. (2017). Predicting Child Labor in Peru: A comparison of Logistic Regression and Neural Networks Techniques. SIMBig. http://ceur-ws.org/Vol-2029/paper5.pdf

Sandri, M., y Zuccolotto, P. (2008). A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees. Journal of Computational and Graphical Statistics, 17(3), 611-628. Recuperado 2 de enero, 2021. http://www.jstor.org/stable/27594328

Walker, S. y Duncan, D. (1967). Estimation of the probability of an event as a function of several independent variables. Biometrika, 167–178.

Wei-Yin, L. (2014). Fifty Years of Classification and Regression Trees. International Statistical Review, 329-348. Recuperado 6 de enero, 2021. http://www.jstor.org/stable/43298996

Zaki, M. y Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms (1ra. ed.)*. Cambridge University Press.

Rokach, L. y Maimon, O. (2014). *Data Mining with Decision Trees - Theory and Applications (2da. ed.)*. Series in Machine Perception and Artificial Intelligence.

Therneau, T. y Atkinson, E. (2019). *An Introduction to Recursive Partitioning Using the RPART Routines*. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Torres, A. (2010). Curvas ROC para Datos de Supervivencia. Aplicación a Datos Biomédicos. Universidad de Santiago de Compostela. http://eio.usc.es/pub/mte/descargas/proyectosfinmaster/proyecto_407.pdf

7. ANEXOS

1. Matriz de Consistencia

Tabla 6

Matriz de Consistencia

PROBLEMA	OBJETIVOS	VARIABLES	METODOLOGIA
Pregunta General	**Objetivo General**	- Variable obtenida de la pregunta: La semana pasada ... ¿Estuvo trabajando o realizando alguna tarea en el hogar o fuera de él para obtener algún ingreso? - Sexo del menor Edad en años del menor - Número de miembros en el hogar - Menores de edad (< 18 años) - Menores de edad (>=18 años) - Último nivel aprobado por el menor - Último nivel aprobado por el jefe(a) del hogar - Clasificación Industrial Internacional Uniforme (CIIU) del Padre - Ingreso bruto del hogar medido en soles - Estrato Socio Económico - Nivel de pobreza clasificados: - Dominio Geográfico - Lima Metropolitana - La vivienda es adecuada - La vivienda cuenta con hacinamiento - La vivienda cuenta con servicios higiénicos - La vivienda esta con dependencia económica - Regiones del Perú	**Tipo** - Correlacional **Diseño** - No Experimental y Transversal
- ¿Cuáles son las características asociadas al riesgo de trabajo infantil en el Perú para el 2019?	- Caracterizar el trabajo infantil en el Perú para el 2019.		
Preguntas Específicos	**Objetivos Específicos**		
- ¿Cuáles son los factores más importantes para caracterizar al trabajo infantil en el Perú para el 2019?	- Identificar los factores más importantes para caracterizar al trabajo infantil en el Perú para el 2019.		
- ¿Cuáles son los perfiles asociados a los trabajadores infantiles en el Perú para el 2019?	- Identificar perfiles asociados a los trabajadores infantiles en el Perú para el 2019.		

2. Ocurrencia de variables por antecedentes

Figura 8

Ocurrencia de Factores en Literatura

	Milagros Jazmín Cayo Velásquez	Luz Aliaga,Maria Guabloche y Mirlene Villacorta	Marco Pariguana(GRADE)	Diego C. Rodrigues, David N. Prata, and Michel A. Silva	Christian Fernando Libaque-Saenz Juan Lazo Karla Gabriela Lopez-Yucra Edgardo R. Bravo	UNICEF, INEC	OIT , CEPAL	Cortex y Gil	Grootaert y Patrinos	
	2018	2009	2011	2015	2017	2015	**2018**	2014	2019	
Modelos	LG	LG	LG	DC	NN	LG	LG	LG	LG-Esp	
Variables	**Perú**	**Perú**	**Perú**	**Brazil**	**Perú**	**Ecuador**	**Brazil**	**México**	**Egipto**	**Ocurrencia**
Sexo	1	1	1	1	1	1	1	1	1	9
Edad	1	1	1		1	1	1	1	1	8
Área de residencia	1	1	1	1	1		1	1	1	8
Nivel educativo de los padres	1	1	1	1	1	1	1	1		8
Estrato Económico		1			1		1	1	1	5
Etnia		1	1				1			3
Asistencia educativa		1			1	1				3
Número de personas en la familia							1		1	2
Tipo de ocupación de los padres						1	1			2
Nro Menores en el Hogar					1					1
Cabeza de Hogar?					1					1
Hermano/a mayor?					1					1
Apoyo del Gobierno en Educación				1						1
Tipo de Vivienda					1					1
Tenencia de Casa					1					1
Edad de los padres								1		1
Atraso Escolar					1					1
Razon #mayores/#menores					1					1
Razon #centros edu /#colegios					1					1

3. Análisis Descriptivo

Tabla 7

Distribuciones menores en las regiones según su situación laboral

Regiones	Trabaja		No Trabaja		Casos
	n	%	n	%	
Cajamarca	330	28.9	810	71.1	1140
Huancavelica	239	26.1	678	73.9	917
Apurímac	185	24.4	574	75.6	759
Huánuco	270	23.4	882	76.6	1152
Áncash	265	22.8	896	77.2	1161
Cusco	236	21.9	842	78.1	1078
Pasco	157	20.5	608	79.5	765
Amazonas	235	19.6	965	80.4	1200
Puno	145	17.4	688	82.6	833
La Libertad	207	15.4	1140	84.6	1347
Junín	154	12.4	1092	87.6	1246
Lambayeque	141	12.3	1009	87.7	1150

Regiones	Trabaja		No Trabaja		Casos
	n	%	n	%	
San Martín	141	11.2	1118	88.8	1259
Ucayali	139	10.3	1213	89.7	1352
Ayacucho	93	9.8	858	90.2	951
Loreto	170	8.3	1873	91.7	2043
Piura	120	7.8	1424	92.2	1544
Madre de Dios	29	5	546	95	575
Tacna	29	4	694	96	723
Moquegua	15	3.3	444	96.7	459
Arequipa	27	2.7	970	97.3	997
Tumbes	10	1.5	645	98.5	655
Lima	39	1.4	2725	98.6	2764
Ica	13	1.3	1000	98.7	1013
Callao	4	0.6	614	99.4	618

4. Dominios Geográficos según INEI

Figura 9

Dominios Territoriales

Nota. El gráfico fue extraído del Ministerio de Vivienda, Construcción y Saneamiento (MVCS)

5. Nodos Terminales

Los siguientes tres nodos corresponden a los menores que se encuentran trabajando.

- Nodo 64: Los menores con un nivel Primaria incompleta menores o iguales a 11 años de las regiones de Áncash, Apurímac, Cajamarca y Huancavelica de los dominios geográficos de la costa norte, sierra y selva provenientes de hogares con jefes(as) en la actividad de agricultura, silvicultura y pesca (3% del total de menores).
- Nodo 66: Los menores con un nivel Primaria incompleta de 10 a 11 años de las regiones de Áncash, Apurímac, Cajamarca y Huancavelica de hogares con jefes(as) en la actividad de agricultura, silvicultura y pesca (1.7% del total de menores).
- Nodo 134: Los menores con un nivel Primaria incompleta menores a 9 años de las regiones de Cusco y Pasco de hogares con jefes(as) en la actividad de agricultura, silvicultura y pesca (0.8% del total de menores).

A su vez se obtuvo 6 perfiles asociados a las características de los que no son trabajadores infantiles.

- Nodo 65: Costa Sur y Lima Metropolitana, de Ancash,4 Apurímac, Cajamarca y Huancavelica, menor o igual a 11 años y provienen de hogares con jefes(as) de hogares realizando actividades de Agrarias, de silvicultura y pesca y posee primaria incompleta (0.2% del total)
- Nodo 135: No son de Cusco, ni de Pasco ni de Ancash, Apurímac, Cajamarca y Huancavelica, menor o igual a 9 años y provienen de hogares con jefes(as) de hogares realizando actividades de Agrarias, de silvicultura y pesca y posee primaria incompleta. (1.4% del total)
- Nodo 17: Poseen primaria incompleta, son de todas las regiones del nodo 2 y provienen de hogares con jefes(as) de hogares realizando actividades que no son Agrarias, de silvicultura y pesca (1.1% del total)
- Nodo 5: No provienen de Amazonas, Ancash, Apurímac, Cajamarca, Cusco, Huancavelica, Huánuco, Pasco ni Puno y su último nivel de estudios es Primaria incompleta. (27.9% del total)
- Nodo 9: No provienen de Amazonas, Ancash, Apurímac, Cajamarca, Cusco, Huancavelica, Huánuco, Pasco ni Puno, su último nivel de estudios es Primaria incompleta y provienen de hogares con jefes(as) de hogares realizando actividades que no son Agrarias, de silvicultura y pesca (4.6% del total)

- Nodo 3: Los menores que no posee como último nivel aprobado Primaria Incompleta (59.4% del total)

6. Recomendaciones

- Al encontrarse como principal variable de importancia las regiones del Perú, nos muestra las diferencias existentes que existen a lo largo del territorio peruano y al ser Perú de los países con mayor presencia de trabajo infantil, se tiene un gran reto de implementar esfuerzos económicos, de salud y educación sobre las regiones con mayor presencia de trabajadores infantiles, además de centrarse en la población descrita por los perfiles encontrados en este trabajo de investigación.
- Se sugiere generar similar información extraída en la ENAHO con un nivel de inferencia no solo departamental si no también distrital para realizar la caracterización territorial del trabajo infantil y este ayude a obtener indicadores en cara a los objetivos de desarrollo sostenible. Por otro lado, se sugiere mejorar el sistema y la transparencia de los datos censales que podrían haber apoyado en la generación de indicadores departamentales.
- Con el árbol de clasificación se obtuvo una tasa de aciertos buena sin incurrir en la complejidad, ahora bien, se recomienda para futuras investigaciones la integración de métodos de balanceo debido al pequeño porcentaje de la clase de interés lo cual podría aumentar la sensibilidad, asimismo el uso de metamodelos basados en árboles de decisión podría ayudar a la mejora de la capacidad predictiva de los trabajadores infantiles.